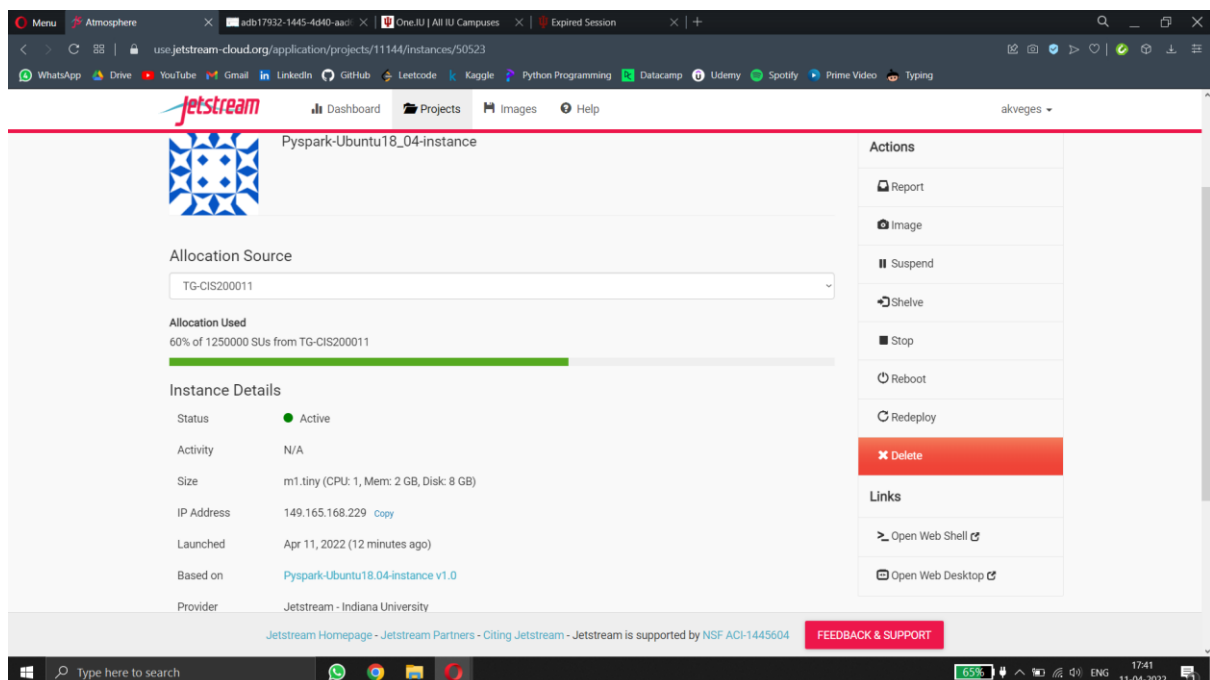


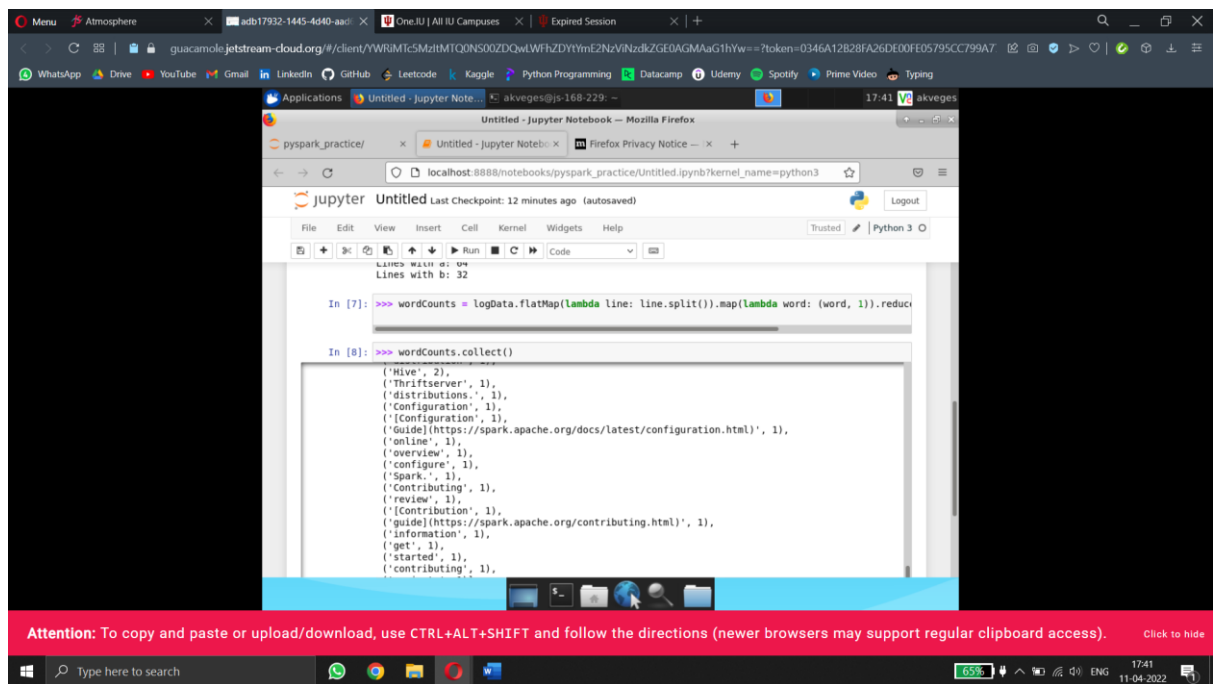
After this week's assignment, I gained a lot of familiarity with Spark and Pyspark. First, we have created a VM instance and selected an image that has all the setups needed to work on. After launching VM, we made Spark installation and made sure the Java, scala, and python installations were successful. Then we installed jupyter notebook py4j. In the notebook, we practiced Pyspark commands. After practicing with the given commands we had to choose a data file and work on it.

I have selected the Peter Pan by J.M. Barrie ebook. I decided to work on it because back when I was a child I used to watch Peter Pan plays. The fictional stories were captivating back when I watched them and I thought it would be really interesting to analyze the text in the ebook and see what are the most and least common meaningful words. Working on this approach RDD and MapReduce will help us to address the 3V's of big data. It will help us perform parallel processing which will help us solve the Velocity and in this way, we can work on massive sizes of data. On the other hand, this helps us work with different kinds of data.

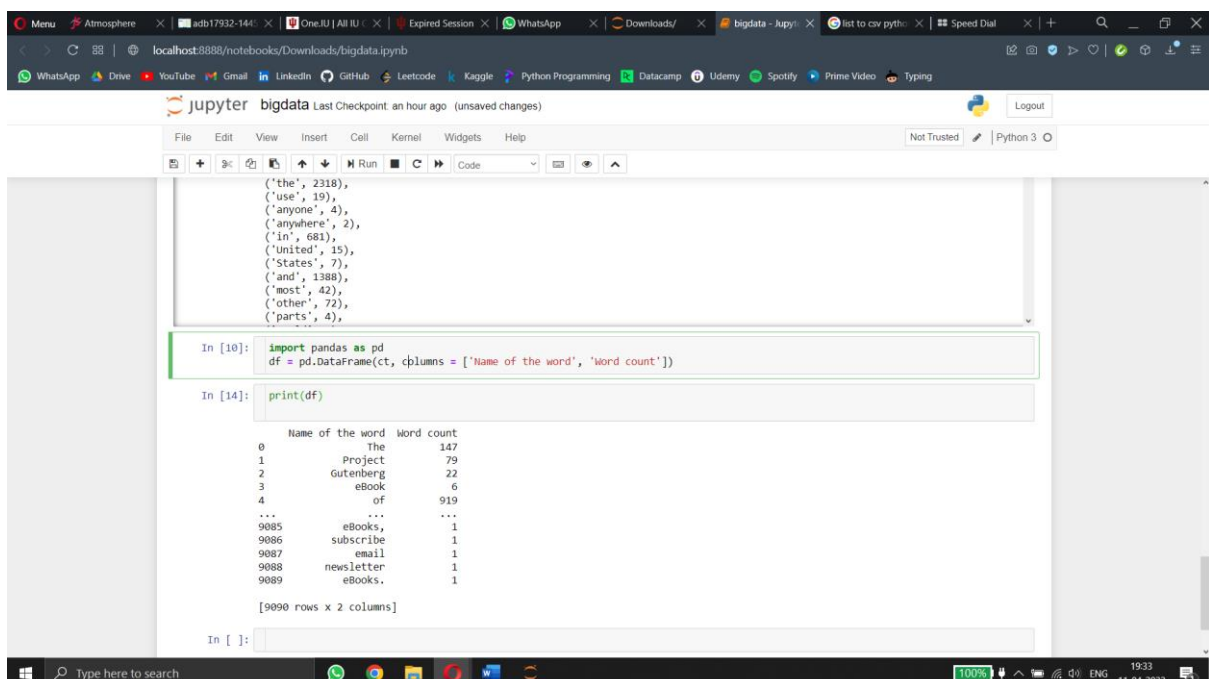
Below is after setting up VM instance



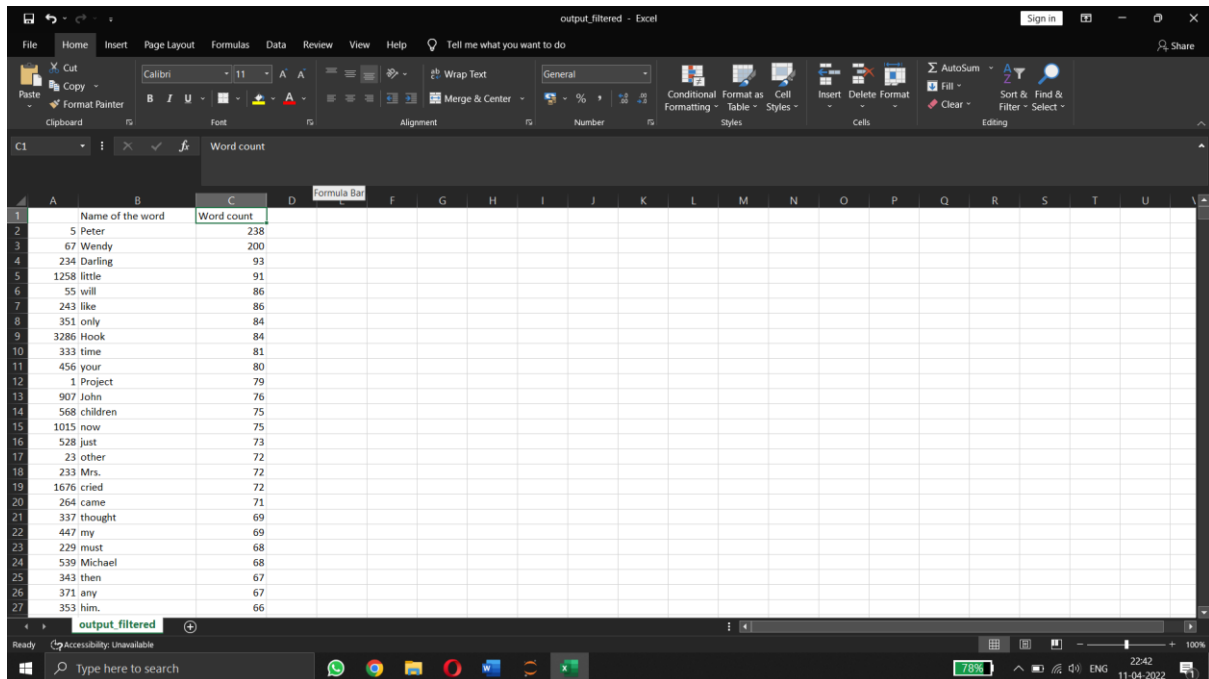
Below is after working on Mapreduce and analysing data with lines of code provided by simple practive exercise in the instructions. It shows the word count list made.



We practiced several RDD objects on the dataset and applied Mapreduce transformations. To work with pandas and other libraries, we loaded the list to pickle in VM instance and unloaded it to local system as a list. Below shows the screenshot of after converting it to a pandas dataframe to work on.



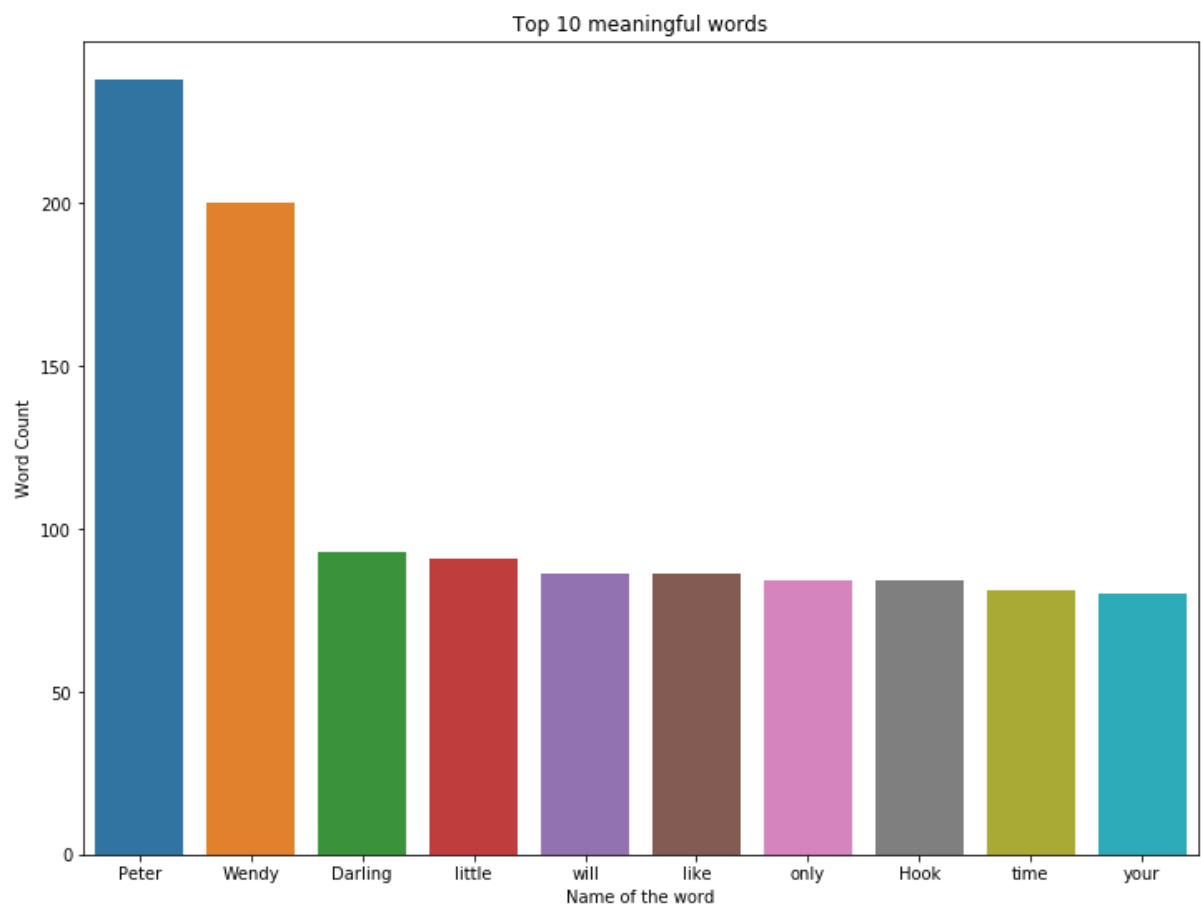
Below shows the data when converted from data frame to csv format, where I filtered out stop words manually. Later after sorting and cleaning the data I imported back to notebook in python as pandas dataframe to create visualizations.

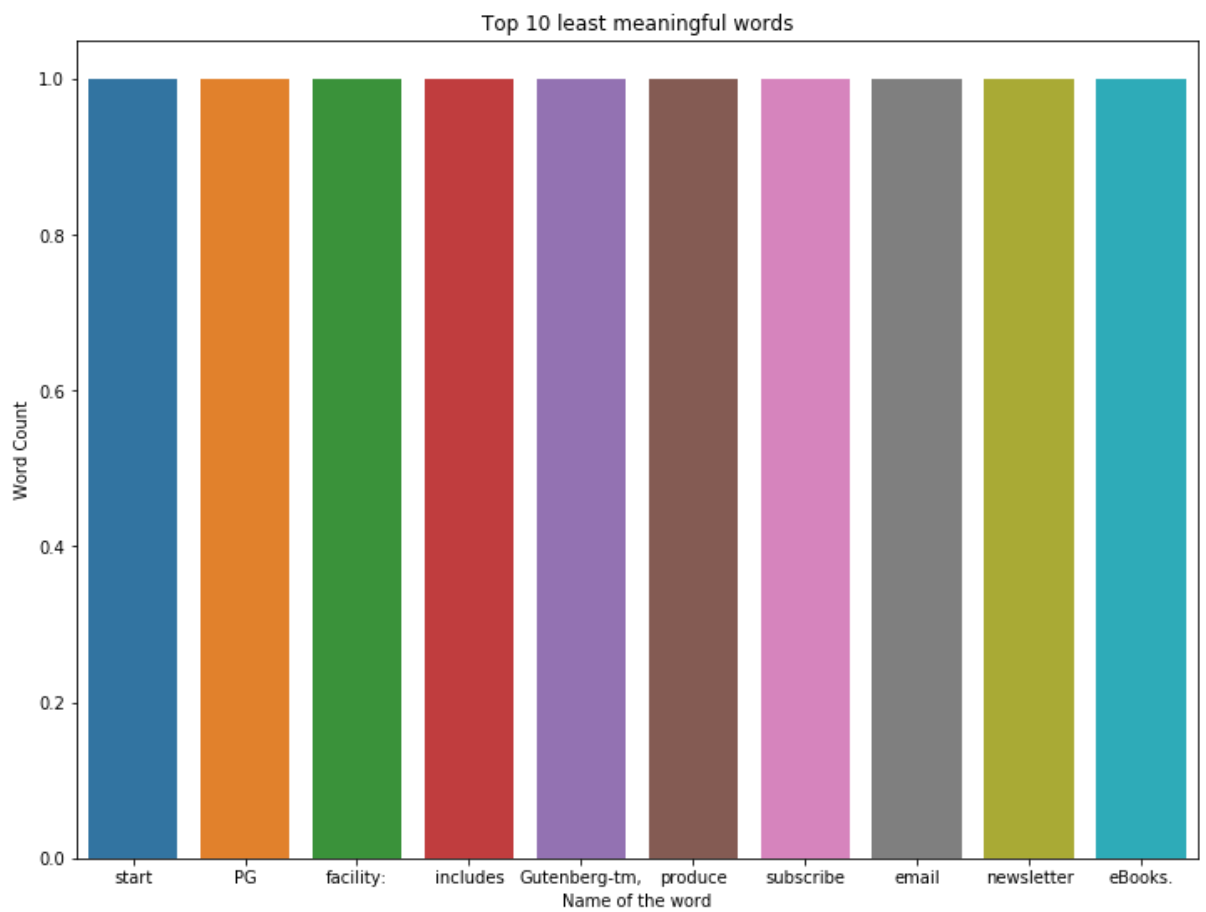


	A	B	C	D
		Name of the word	Word count	Formula Bar
1				
2	5	Peter	238	
3	67	Wendy	200	
4	234	Darling	93	
5	1258	little	91	
6	55	will	86	
7	243	like	86	
8	351	only	84	
9	3286	Hook	84	
10	333	time	81	
11	456	your	80	
12	1	Project	79	
13	907	John	76	
14	568	children	75	
15	1015	now	75	
16	528	just	73	
17	23	other	72	
18	233	Mrs.	72	
19	1676	cried	72	
20	264	came	71	
21	337	thought	69	
22	447	my	69	
23	229	must	68	
24	539	Michael	68	
25	343	them	67	
26	371	any	67	
27	353	him	66	

So using matplotlib and seaborn libraries I have created barplots showing top 10 and least 10 meaningful words in the bag of words and plotted them against the word count. This will show

the top and bottom 10 meaningful words. Below are the bar plots .





The barplots show the most common meaningful work is Peter and the least is ebooks. The subsequent 10 are shown in the plots.